

Parallel calculation of the median and order statistics on GPUs with application to robust regression

Gleb Beliakov

School of Information Technology,

Deakin University,

221 Burwood Hwy, Burwood 3125, Australia

gleb@deakin.edu.au

Abstract

We present and compare various approaches to a classical selection problem on Graphics Processing Units (GPUs). The selection problem consists in selecting the k -th smallest element from an array of size n , called k -th order statistic. We focus on calculating the median of a sample, the $n/2$ -th order statistic. We introduce a new method based on minimization of a convex function, and show its numerical superiority when calculating the order statistics of very large arrays on GPUs. We outline an application of this approach to efficient estimation of model parameters in high breakdown robust regression.

Keywords GPU, median, selection, order statistic, robust regression, cutting plane.

I. INTRODUCTION

We consider a classical selection problem: calculate the k -order statistic of a sample $\mathbf{x} \in \mathbb{R}^n$, $x_{(k)}$. Calculation of the median of a sample is a special instance of the selection problem, with $Med(\mathbf{x}) = x_{((n+1)/2)}$, where $[t]$ denotes the integer part of t . Selection problem has various applications in data analysis, in particular in high-breakdown regression [19], [27], which is a motivating application discussed later in this paper. Selection problem is also valuable for other high-performance computing applications, such as image processing, computational aerodynamics and simulations, and it has received attention in parallel computing literature [1], [2], [30], [34].

Our concern is efficient calculation of the order statistics of very large vectors by offloading computations to GPUs. General purpose GPUs have recently become a powerful alternative to traditional CPUs, which allow one to offload various repetitive calculations to the GPU *device*. GPUs have many processor cores ($p = 240$ in NVIDIA's Tesla C1060 and $p = 448$ in C2050), and can execute thousands of threads concurrently [23]. The computational paradigm here is SIMT (single-instruction multiple-thread). Different streaming multiprocessors can execute different instructions, although threads within groups (so called *warps*) are executed in SIMD (single instruction multiple data) fashion. The peak performance is achieved when there are no, or almost no branching due to conditional statements.

The selection problem is related to sorting. A naive approach to the selection problem is to sort the elements of \mathbf{x} and select the desired order statistic. On CPU, the complexity of the sort

operation is $O(n \log n)$ by using one of the standard sorting algorithms. Radix sort has $O(nb)$ complexity where b is the number of bits in the key [31].

An improvement over sorting is to use a selection algorithm, such as *quickselect* [24], [31], which is based on *quicksort*, and has an expected complexity of $O(n)$. The median of medians algorithm developed in [4], also known as *BFPRT* after the last names of the authors, has $O(n)$ worst case complexity, although empirical evidence suggests that on average *quickselect* is faster.

The sort operation was parallelized for GPUs. Parallel sorting algorithms include GPU quicksort [8], GPU sample sort [18], bucket-merge hybrid sort [33], GPU radix sort [20], [29] and others [12], [14], [32]. At the time of writing, GPU radix sort, which was implemented in Trust software [15], and in the CUDA 4.0 library [22] released in March 2011, is the most efficient GPU sort algorithm. In empirical tests it delivered performance of up to 580 million elements/sec (for 32 bit keys on Tesla C2050) [29], which was consistent with our own evaluation. There have been attempts to parallelize selection algorithms [1], [2], [30], [34] for coarse-grained multicomputers, and in the the framework of bulk-synchronous parallelism paradigm, but these methods use divergent threads of computation. To our knowledge there is no parallel version of a selection algorithm suitable for GPUs at the moment.

In this contribution we analyze various alternatives for parallel calculation of the median of a large real vector ($n > 10^5$) on a GPU device, and propose a new parallel selection method. Rather than attempting to parallelize an existing serial selection algorithm, we propose an alternative method, based on minimization of a convex cost function. Calculation of the cost function is performed by reduction in parallel. By using an effective minimization algorithm, which requires only a few iterations, we design a cost efficient approach to parallel selection.

The paper is structured as follows. In Section II we outline the approaches to calculating the median and order statistics in parallel and list our alternatives suitable for GPUs. In Section III we present a new approach based on minimizing a convex function, or alternatively, on solving a nonlinear equation. Section IV presents Kelley's cutting plane algorithm, which we found to be the most efficient and stable among the optimization methods we evaluated. Section V is devoted to comparison and benchmarking of different alternatives for selection on GPUs. In Section VI we outline two applications that require efficient selection algorithms, taken from the area of data analysis. Section VII concludes.

II. APPROACHES TO SELECTION ON GPU

Consider a vector $\mathbf{x} \in \mathbb{R}^n$, n of order of $10^5 - 10^9$. We are interested in efficient calculation of its median, or an order statistic. As we mentioned in the Introduction, this can be achieved by using either sorting or selection algorithms.

Our problem has a specific feature. The vector \mathbf{x} is stored as an array of floats (single or double precision) on the GPU *device*. The reason is that often the components of \mathbf{x} are themselves calculated in parallel on GPU. This is the case in our motivating application.

In addition, we consider a situation where \mathbf{x} is so large that it does not fit the memory of a single GPU device nor CPU memory, and is distributed over several GPU devices.

Let us list our alternatives.

- 1) Perform parallel sorting on GPU by any available GPU sorting algorithm. In our study we used the fastest GPU radix sort (we benchmarked several GPU sorting algorithms, and the results are consistent with the conclusions by Merrill and Grimshaw [20]).

- 2) Transfer \mathbf{x} to CPU and use *quickselect* algorithm.
- 3) Use *quickselect* on GPU running as a single thread.
- 4) Develop an alternative GPU selection algorithm.

We note that when a large number of calculations of order statistics (in particular, the medians) of different vectors are required, even small gains in computational efficiency become important.

All the listed alternatives seem to be feasible and competitive. The cost of alternative 2 includes the cost of copying data between GPU and CPU. The *quickselect* algorithm is very efficient on modern CPUs, as compared to other sorting and selection methods. Alternative 3 eliminates the need to transfer the data, but at the expense of much slower execution of *quickselect* on one of the GPU cores.

When the data are distributed over several GPUs, alternatives 2 and 3 are no longer feasible, and GPU sorting algorithms may require substantial modifications and will degrade in performance, since the data needs to be moved between different GPU devices. However, the nature of our alternative selection algorithm makes it applicable to multiple GPU situation.

III. APPROACH BASED ON SEARCH AND OPTIMIZATION

It was known to Laplace and Gauss that the arithmetic mean, the median, the mode, and some other functions are solutions to simple optimization problems [5], [11], [16]. A recent account and generalizations can be found in [6], [7], [35]. In particular, the median is a minimizer of the following expression:

$$Med(\mathbf{x}) = \arg \min_y f(y) = \arg \min_y \sum_{i=1}^n |x_i - y|. \quad (1)$$

The objective is clearly non-smooth (it is piecewise linear), but it is convex, because the sum of convex functions is always convex. Convexity of the objective implies that there exists a unique minimum of the above expression, which can be easily found by a number of classical univariate optimization methods. Of course, numerical solution to problem (1) is more expensive than *quickselect* on CPUs, but taking into account that the data reside on the GPU, and that each evaluation of all the terms in the objective can be done in parallel in $O(\frac{n}{p})$ operations (p is the number of processor cores), this method becomes a promising alternative.

For calculation of a k -th order statistic, the following objective is used [6], [7]

$$OS_k(\mathbf{x}) = \arg \min_y \sum_{i=1}^n u(x_i - y), \quad (2)$$

where

$$u(t) = \begin{cases} (n - k + \frac{1}{2})t & \text{if } t \geq 0, \\ -(k - \frac{1}{2})t & \text{if } t < 0. \end{cases}$$

The objective (2) is also piecewise linear convex.

For the purposes of optimization, we considered the following alternatives: 1) Brent's method [24], 2) the golden section method, 3) an analog of the Newton's method for non-smooth functions [3], and 4) Kelley's cutting plane method [17].

When the objective f is differentiable, a minimization problem can be converted into a root-finding problem: solve $f'(y) = 0$. For non-smooth functions we need to solve the following nonlinear equation

$$0 \in \partial f(y),$$

where ∂f is the Clarke's subdifferential [9], [10]. For a univariate function, Clarke's subdifferential at a point t is simply a set of values $\partial f(t) = \{z \in \mathbb{R} \mid \forall y \in \mathbb{R}, f(y) \geq z(y - t) + f(t)\}$, such that any line with the slope $z \in \partial f(t)$ which passes through $(t, f(t))$, is a tangent line to the graph of f . An element of the subdifferential is called a subgradient.

We know that for the modulus function

$$\partial|t| = \begin{cases} \{-1\} & \text{if } t < 0, \\ [-1, 1] & \text{if } t = 0, \\ \{1\} & \text{if } t > 0. \end{cases}$$

We can express ∂f using $g : \mathbb{R} \rightarrow I \subseteq \mathbb{R}$, where I is a set of intervals on a real line,

$$g(y) = \text{count}(x_i > y) + \text{count}(x_i < y)(-1) + \text{count}(x_i = y)[-1, 1],$$

where *count* is the number of elements of \mathbf{x} satisfying a given condition, and where the usual Minkowski sum for sets is used.

There are many root-finding algorithms, which can all be adapted to equation $0 \in g(y)$. In particular, we adapted the classical bisection method, similar to how it was done in [13] but in a different context, as well as the method of parabolas combined with golden section, which is also known as Brent algorithm [24]. In fact this root finding algorithm is equivalent to Brent's optimization method, however we looked at both alternatives, as implementation details and stopping criteria may play a role in their efficiency.

Evaluation of the objective in (1) requires summation. On GPUs it can be performed in parallel by *reduction*, using binary trees in $O(\frac{n}{p} + \log p)$ operations. Reduction on GPU is efficiently implemented and relies on various hardware related aspects and strategies, like loop unrolling [21]. Furthermore, calculation of (1) on GPU involves only reading from (slow) global memory in coalescent blocks, but no writing when swapping the elements of \mathbf{x} , as in sorting algorithms (all temporary results are written into a much faster shared memory). As for Eq. (2), used for order statistics, it introduces only minimal branching in calculating each term of the sum. In the experiments we conducted, this did not produce any noticeable effect on the run time of the algorithm.

IV. KELLEY'S CUTTING PLANE METHOD

The methods of function minimization and root finding we have used are classical, and are discussed in detail in several textbooks, e.g. [24]. Kelley's cutting plane method [17] is less known. In addition, as we show in the subsequent sections, it was the most stable and efficient algorithm in our study. Therefore it is worth to outline Kelley's algorithm in this section.

The cutting plane method is relying heavily on the convexity of the objective, and requires calculation of both the objective f and its subgradient g . This method iteratively builds a lower piecewise linear approximation to the objective.

At each iteration, the algorithm evaluates the value of the objective f and its subgradient g . A piecewise linear lower estimate of f is built using the formula

$$h(y) = \max_{i=1,\dots,k} \{g(y_i)(y - y_i) + f(y_i)\},$$

where y_i are the points the objective was evaluated during the previous k iterations. The minimizer of h is chosen as the y_{k+1} . The algorithm starts with $y_1 = x_{(1)}$ and $y_2 = x_{(n)}$, which are evaluated on GPU in parallel using reduction. Figure 4 illustrates these steps.

It is worth noting that the minimizer of h is always bracketed by two values y_L and y_R , initially y_1 and y_2 , and is found explicitly by

$$y_{k+1} = \frac{f(y_R) - f(y_L) + y_L g(y_L) - y_R g(y_R)}{g(y_L) - g(y_R)}.$$

y_L and y_R are updated in the following way: if $g(y_{k+1}) < 0$ then $y_L \leftarrow y_{k+1}$ and if $g(y_{k+1}) > 0$ then $y_R \leftarrow y_{k+1}$. If $g(y_{k+1}) = 0$, the solution is found.

Let us present Kelley's cutting plane algorithm for completeness.

Algorithm 1 (Cutting plane algorithm): Input: f - the objective function, g - a subgradient of f , $maxit$ - the upper bound on the number of iterations.

Input/output: y_L, y_R - the ends of the interval containing the minimizer.

Output: y - minimizer found.

0. $fL \leftarrow f(y_L), gL \leftarrow g(y_L), fR \leftarrow f(y_R), gR \leftarrow g(y_R)$.
1. for $i = 1; i \leq maxit; i = i + 1$ do:
 - 1.1 $t \leftarrow (fR - fL + y_L * gL - y_R * gR) / (gL - gR)$.
 - 1.2 $ft \leftarrow f(t), gt \leftarrow g(t)$.
 - 1.3 If Stopping criteria then $y \leftarrow t$ and exit.
 - 1.4 If $gt < 0$
 - then $y_L \leftarrow t, fL \leftarrow ft, gL \leftarrow gt$.
 - else $y_R \leftarrow t, fR \leftarrow ft, gR \leftarrow gt$.
2. $y \leftarrow t$ and exit.

The stopping criteria at step 1.3 comprise the following: $gt = 0$ (the point with $0 \in \partial f(t)$ was found), $y_R - y_L \leq tolerance_f$, $|gt| \leq tolerance_g$.

The cutting plane algorithm in one dimension has little overhead, and both f and g are easily computed in parallel and simultaneously on a GPU device. Once an approximation \tilde{y} to the median is computed with the desired accuracy, a simple loop finds the exact value of the median in $O(\frac{n}{p} + \log p)$ operations on a GPU device by reduction¹.

The cutting plane algorithm starts with calculation of f and g at the extremes of the range of the data (at step 0). The values $y_L = y_1 = x_{(1)}$ and $y_R = y_2 = x_{(n)}$ can be calculated in parallel by reduction. Let us also notice that the values of f and g at these points are computed by simple formulas: $g(y_L) = -n + 2$, $g(y_R) = n - 2$, $f(y_L) = \sum x_i - ny_L$ and $f(y_R) = ny_R - \sum x_i$. The values y_L, y_R and $\sum x_i$ can be computed in a single parallel reduction operation, which is advantageous compared to computations of $y_L, y_R, f(y_L)$ and $f(y_R)$ independently using four reductions. Therefore the complexity of the Algorithm 1 is at most $maxit + 1$ parallel reductions.

¹This is a simple loop which selects the largest element $x_i \leq \tilde{y}$.

To end this section, we notice that the cutting plane method takes only a few iterations to converge to an approximate solution \tilde{y} , under 30 iterations in our experiments with n up to 32 million and $tolerance_f = 10^{-12}$. Furthermore, we improved the runtime of the algorithm by combining it with parallel sorting as follows. First, we run the cutting plane method for very few iterations, say 5-7 iterations. After completing these iterations, we have an interval $[y_L, y_R]$ containing the minimizer, which is the median of \mathbf{x} . We think of this interval as a pivot in selection algorithms. At the second stage, we select the values x_i which fall within the (open) pivot interval $]y_L, y_R[$ and copy them to a smaller array \mathbf{z} . This can be done in parallel on GPU using *copy_if* operation. Then the elements of \mathbf{z} are sorted in parallel using GPU radix sort. The median is then the k -th order statistic $z_{(k)}$ with $k = \lfloor \frac{n}{2} \rfloor - m$, and m being the number of elements of \mathbf{x} smaller than or equal to y_L , recorded during execution of the cutting plane method.

This way we obtain a hybrid selection algorithm, which benefits from the fact that sorting is performed on a much smaller array \mathbf{z} . The number of iterations of the cutting plane method is selected to maximize efficiency: the algorithm stops when the cost of its iteration outweighs the benefit of further reducing the pivot interval for faster conditional copy and sorting. This number is empirically selected for a given architecture. In our experiments we stopped the cutting plane algorithm after 7 iterations (for $n = 2^{25}$), as at that time the pivot interval contained under 2^{19} elements, and its sorting was already very fast.

V. EMPIRICAL STUDY OF COMPUTATIONAL EFFICIENCY

A. Data sets

We selected the following methods for numerical comparison: *quickselect* on CPU, *quicksselect* on GPU as a single thread, GPU version of the radix sort [29], four methods based on minimization of (1) (Brent's, golden section, nonsmooth quasi-Newton and cutting plane) and two methods based on solving $0 \in g(y)$ (bisection and Brent's root finding algorithm).

We randomly generated the following data sets of varying length $n \in \{8192 = 2^{13}, 32768 = 2^{15}, 131072 = 2^{17}, 524288 = 2^{19}, 2097152 = 2^{21}, 8388608 = 2^{23}, 33554432 = 2^{25}, 134 \times 10^6 \approx 2^{27}\}$:

- 1) Uniform $x_i \sim U(0, 1)$
- 2) Normal $x_i \sim N(0, 1)$
- 3) Half-normal $x_i = |y_i|$ and $y_i \sim N(0, 1)$
- 4) Beta $x_i \sim \beta(2, 5)$
- 5) Mixture 1, 66.6% of elements of x_i chosen from $N(0, 1)$ and 33.3% from $N(100, 1)$
- 6) Mixture 2 50% of elements of $x_i + 1$ chosen from $N(0, 1)$ and the rest from $N(100, 1)$
- 7) Mixture 3 90% of elements of x_i chosen from half-normal $N(0, 1)$ and the rest set to 10.
- 8) Mixture 4 66.6% of elements of x_i chosen from half-normal $N(0, 1)$ and 33.3% from $N(100, 1)$
- 9) Mixture 5 50% of elements of $x_i + 1$ chosen from half-normal $N(0, 1)$ and the rest from $N(100, 1)$

In addition, we performed experiments in which one or more components of \mathbf{x} took very large values $\sim 10^9$. The reasons for our selection of the distributions are the following. The median is scale invariant, so the parameters of U and N are taken without loss of generality. We wanted

to test the algorithms for symmetric and asymmetric distributions, such as half-normal and beta, as well as for mixtures. The use of half-normal distributions is motivated by our background application, that of regression, which is described in the next section. If the data follow a linear model with normally distributed noise, the absolute residuals follow a half-normal distribution. If there are large outliers in the data, we will have a mixture of distributions for the residuals. We tested mixtures with both equal and unequal proportions of the components. These proportions correspond to the proportion of the outliers in our background application. The model with a mixture of half-normal and normal is the closest to our application, and in fact this distribution was observed when we used regression residuals as the data.

B. Algorithms and testing environment

Our preliminary analysis allowed us to exclude several algorithms. Golden section was inferior to Brent’s method. In fact, Brent’s algorithm from [24] is organized in such a way that it uses the method of parabolas, and reverts to the golden section if the method of parabolas is unsuccessful, so it is always no worse than golden section. The quasi-Newton method was very unstable, and failed to converge in most cases.

Thus we performed detailed numerical comparison of seven algorithms: *quickselect*, *quickselect* on GPU, GPU radix sort, Brent’s method of optimization, cutting plane, bisection and Brent’s method of solving nonlinear equations.

We measured the average time of each algorithm over 10 instances of each data set for every fixed n , and repeated calculations for each instance 10 times. We excluded the time spent on loading the data set from a file and on transferring it to the GPU device.

Our hardware was Tesla C2050 with $p = 448$ cores and 3 GB of RAM, connected to a four-core Intel i7 CPU with 4 GB RAM clocked at 2.8 GHz, running Linux (Fedora 12). Before presenting the results, we mention some of the key performance parameters. Transfer of a 32M array of floats/doubles from GPU to CPU on our system takes over 230/455 ms, while transfer of a 500K array takes only 4/6.1 ms. Radix sort on GPU takes 65ms for a 32M array of floats, but degrades to 226 ms for 32M array of doubles, which are similar values to those reported in [29]. One reduction on GPU when calculating objective (1) or its subgradient g took 3ms for a 32M array of doubles (1.9 ms for floats).

In our implementation of the algorithms, we relied on Thrust library [15], in particular, Thrust’s implementation of radix sort, and the *transform_reduce* and *copy_if* functions. Thrust library offers a number of high-level algorithms, such as reduction and transformations with arbitrary unary and binary operators, and automatically selects the most appropriate configuration parameters (such as block size and number of blocks in a warp). Therefore we do not report experiments with these parameters, assuming that the optimal choice was made by Thrust.

We outline the program code used to calculate the the objective f and its subgradient g needed by Algorithm 1 in Figure 1. The actual implementation of our algorithm called `cp_select` library, can be downloaded from http://www.deakin.edu.au/~gleb/cp_select.html. We want to stress the simplicity of the algorithm, which takes only a few lines of code relying on a high-level interface to standard algorithms, such as reduction and conditional copy, implemented in the Thrust library.

C. Empirical results

We present the results of our numerical experiments in Tables I and II for single and double precision respectively, and graphically in Figures 2,3 (the plot is the log-log plot). We present the average results for all distributions, because we did not observe any significant deviations from the average values for different distributions of the data. In fact, for the most efficient GPU radix sort and the proposed method, CPU times were very consistent: the largest and the smallest CPU time were within less than 5% from the average values. For *quickselect* (both CPU and GPU versions) the variations were larger, but we did not observe any pattern related to the type of the distribution, only variations between the instances of data from the same distribution.

We observe that the use of *quickselect* on the CPU was more efficient only for the smallest data set tested $n = 8192$, and that for larger data sets GPU radix sort was faster by a factor of thirteen and five for single and double precision respectively. Both copying the data and actual selection algorithm on CPU have contributed to the cost of *quickselect* on CPU. We see from the tables that the time spent on just copying the data to CPU was larger than that of GPU radix sort starting from $n \geq 2^{15}$, and the same was true for the time used by *quickselect* algorithm itself. Execution of *quickselect* on GPU in a single thread was worse than the remaining alternatives by a very large factor.

GPU radix sort was the most efficient method for up to $n = 2^{21}$. At that point the methods based on minimization outperformed GPU radix sort and consistently remained more efficient for large data sets. It looks from the plots of the logarithms of CPU time, that starting from $n = 2^{23}$ the CPU time behaved like $O(n)$ for all methods, and the difference was due to a constant factor. The difference in performance of GPU radix sort and the proposed method was more pronounced for double precision values, as one would have expected, because the performance of the radix sort depends on the number of bits in the key.

Among the methods of minimization/root finding, bisection was the slowest. Brent's root finding algorithm delivered similar performance to the cutting plane method, but its performance degraded when data contained very large outliers, as explained below.

The difference in CPU time between the fastest existing alternative GPU radix sort and our cutting plane method is almost six-fold for large arrays $n = 2^{27}$ (double) and three-fold (single precision).

D. Discussion

We note that unlike traditional sorting and selection algorithms, our methods based on minimization and root-finding are not sensitive to how the data are organized in the array \mathbf{x} (e.g., at random, pre-sorted, or sorted in inverse order). Indeed expression (1) is invariant with respect to permutations of the components of \mathbf{x} .

However, the performance of all algorithms based on minimization or solving nonlinear equations, except the cutting plane method, drastically decreased when just one or a few components of \mathbf{x} took very large values (e.g., 10^9), which could happen in our application. We do not report the CPU times, because they depend on how large was the outlier. In fact we could slow down the mentioned algorithms to any degree by taking a sufficiently large outlier. That is, all of our methods except cutting plane, are sensitive to the values of x_i . The reason is the following.

Clearly, the number of iterations of the bisection method required to achieve a fixed accuracy is $O(\log r)$, where $r = x_{(n)} - x_{(1)}$ is the range of the data, which can be unbounded. The same is true for the golden section optimization method, which is the counterpart of the bisection. Both Brent’s methods reverted to the slower bisection/golden section methods because the objective is linear in almost all of the range of the data when x_i have such an uneven distribution. Hence the parabolic fits attempted by Brent’s method are unsuccessful, see Figure 5.

The only optimization method insensitive to very large (or small) values of x_i was Kelley’s cutting plane method. The reason this method was successful is that it uses not only the values of the objective but also its subgradients, as well as convexity of the objective. The cutting plane method iteratively builds a lower piecewise linear approximation to the objective, which also happens to be piecewise linear. In that way, just one value of the objective and its subgradient allows the algorithm to eliminate large uninteresting linear pieces (such as the intervals between a large outlier x_i and the bulk of the data), which do not contain the minimizer.

However, even the cutting plane method had problems when some components of \mathbf{x} were extremely large, of order of 10^{20} , because of finite precision of machine arithmetic. When such a large value is accounted for in the sum (1), all the subsequent terms do not contribute to the sum because of loss of precision, even though their combined value can be sufficiently large. This is a well known effect in numerical analysis.

To circumvent this issue we used the following approach. The order statistics, and the median in particular, are invariant with respect to monotone transformations. Then we can transform the elements of \mathbf{x} by applying an increasing function F componentwise $F(\mathbf{x}) = (F(x_1), \dots, F(x_n))$, calculate the median of the transformed array $med_F = Med(F(\mathbf{x}))$, and then take $med = F^{-1}(med_F)$. In our work we used transformation $F(t) = \log(1 + t - x_{(1)})$. In this case the loss of accuracy in summation is eliminated.

As we mentioned in Section IV, the complexity of the cutting plane algorithm is $maxit + 1$ parallel reductions, plus the time spent on conditional copy and sorting a reduced array \mathbf{z} . In all our experiments, the array \mathbf{z} turned out to be significantly smaller than \mathbf{x} , typically between to 1% and 5% of the length of \mathbf{x} . Even if, hypothetically, \mathbf{z} is almost as large as \mathbf{x} , the complexity of the proposed method is tied to the complexity of the GPU radix sort, because the use of cutting plane adds only a fixed initial overhead. We were unable to design a data set which would lead to such a hypothetical case.

Finally, we will mention the situation where the data are distributed across several GPUs. Parallel sorting algorithms require transfers of data between GPUs, which has a non-negligible cost, even when such transfers do not involve the CPU, and unified virtual addressing is used, as in CUDA 4.0. In contrast, calculation of (1) and its subgradient is embarrassingly parallel, and involves reductions executed independently on different GPUs. The partial sums from several GPUs are added together on the CPU. It works in a similar way for multiple GPUs connected to different hosts, which are in turn connected through MPI interface, as only small portions of data need to be transferred on few occasions (of course, we assume that the number of GPUs is not large). Therefore we see the approach based on minimization as the most suitable alternative to selection on multiple GPUs

VI. APPLICATION TO ROBUST REGRESSION

Next we outline our motivating problem of high breakdown regression, the role of the median function and our GPU based algorithm. We also mention another application from data analysis which requires multiple calculations of the order statistics.

We consider the classical linear regression problem: given a set of pairs $\{(x_i, y_i)\}, i = 1, \dots, n$: $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ (data), and a set of linear models $f_\theta : \mathbb{R}^p \rightarrow \mathbb{R}$ parameterized by a vector of parameters $\theta \in \Omega \subseteq \mathbb{R}^p$ determine the parameter vector θ^* , such that f_{θ^*} fits the data best. The linear models are

$$y_i = x_{i1}\theta_1 + \dots + x_{ip}\theta_p + \varepsilon_i, \quad i = 1, \dots, n, \quad (3)$$

with $x_{ip} = 1$ for regression with an intercept term. $\{x_{ij}\} = X \in \mathbb{R}^{n \times p}$ is the matrix of explanatory variables and ε is an n -vector of iid random errors with zero mean and unknown variance.

The goodness of fit is expressed in terms of either squared or absolute residuals $r_i = f_\theta(x_i) - y_i$, namely the weighted averages $\sum_{i=1}^n w_i r_i^2$ (the least squares (LS) regression), or $\sum_{i=1}^n w_i |r_i|$ (the least absolute deviations (LAD) regression).

The LS and LAD problems are not robust to outliers: it is sufficient to take just one contaminated point with an extremely large or small value of y_i or x_{ij} to break the regression model. The breakdown point of the LS and LAD estimators, i.e., the proportion of data that can make the estimator's bias arbitrarily large, is 0 (see books [19], [27]). To overcome the lack of robustness of the LS and LAD estimators, Rousseeuw [25] introduced the Least Median of Squares (LMS) estimator based on the solution of

$$\text{Minimize } F(\theta) = \text{Med}(r_i(\theta))^2.$$

He also introduced the method of the least trimmed squares (LTS) [25], which is considered superior to the LMS [26], [28]. Here the expression to be minimized is

$$\text{Minimize } F(\theta) = \sum_{i=1}^h (r_{(i)}(\theta))^2,$$

where the residuals are ordered in the increasing order $|r_{(1)}| \leq |r_{(2)}| \leq \dots \leq |r_{(n)}|$, and $h = \lfloor (n + p)/2 \rfloor$.

In essence, in the LMS and LTS methods, half of the sample is discarded as potential outliers, and the model is fitted to the remaining half. This makes the estimator not sensitive to contamination in up to a half of the data. Of course, the problem is to decide which half of the data should be discarded, and this is very challenging.

Numerical computation of the LMS or LTS estimator involves multiple evaluations of $F(\theta)$. We clearly see the need in an efficient evaluation of the median in the LMS method, which is the motivation for this work. Next we also show that the median can simplify calculation of the LTS objective, even though it appears that the vector of residuals needs to be fully sorted.

We rewrite the LTS objective in this form

$$\text{Minimize } F(\theta) = \sum_{i=1}^n \rho(|r_{(i)}(\theta)|^2), \quad (4)$$

with

$$\rho(t) = \begin{cases} 1 & \text{if } t < \text{Med}(\mathbf{r}) \\ \frac{a}{b} & \text{if } t = \text{Med}(\mathbf{r}) \\ 0 & \text{otherwise} \end{cases}$$

The integers a and b are computed from the multiplicity of the median, i.e., the number of components of \mathbf{r} equal to the median. Consider the numbers $b_L = \text{count}(|r_i| < \text{Med}(\mathbf{r}))$ and $b = \text{count}(|r_i| = \text{Med}(\mathbf{r}))$. In the LTS method take $h = \frac{n+1}{2}$ for odd n , and $h = \frac{n}{2}$ for even n . Then $h = b_L + a$ for some $a \leq b$. Now if we take a and b calculated in this way in (4), the value of F in (4) will be the sum of exactly h smallest residuals, and will coincide with F in the LTS method. The values of a and b are easily calculated once the median is known, on either CPU (in $O(n)$ time) or GPU (in $O(\frac{n}{p} + \log p)$ time) by reduction.

Hence, calculation of the median is also valuable for the LTS method, where partial sorting of the absolute residuals can be replaced with a cheaper method based on the median.

Another application which requires multiple calculations of order statistics, and which is easily parallelized for GPUs is the k -nearest neighbor method (kNN), which is used in both regression and classification. In the case of regression, it consists in approximating a function value at \mathbf{x} by $f(\mathbf{x}) = \sum_{i=1}^k w_i f_i$, where the values f_i are the ordinates of the k points $\mathbf{x}_1, \dots, \mathbf{x}_k$ closest to \mathbf{x} (in some metric, usually the Euclidean distance) and w_i are the weights that are decreasing functions of the distances $d_i = \|\mathbf{x} - \mathbf{x}_i\|$. In the case of classification, the majority vote (among the k nearest neighbors) is applied to determine the predicted class of the query point \mathbf{x} .

The kNN method is suitable for parallelization on GPUs. The array of distances \mathbf{d} , can be calculated on GPUs in parallel in $O(\frac{n}{p})$ time. The usual approach to selecting the k nearest neighbors is to sort the data according to the distance to \mathbf{x} . However, one can do better by using the k -th order statistic. Indeed, by adapting the function ρ in (4), we obtain an indicator function, which returns a non-zero value for those data (\mathbf{x}_i, f_i) that are no further from \mathbf{x} than the k -order statistic $d_{(k)}$. Then the weighted sum of k nearest neighbors is calculated by reduction. Thus, efficient parallel calculations of k -order statistics is useful in the kNN method as well.

VII. CONCLUSION

We considered a classical selection problem and discussed various approaches to parallel computation of the median and order statistics on GPUs. We compared numerical performance of several existing and proposed methods by using a number of data sets with standard and unusual distributions, and found that the most efficient way to calculate the median is by minimizing a convex function using the cutting plane method, followed by sorting a reduced sample. This approach is between three and six times more efficient than the fastest existing method based on GPU radix sort, depending on the data type used.

The proposed algorithm is easy to implement using high-level interface to GPU programming implemented in Thrust library, which is bundled with the newest CUDA 4.0 release. It uses a standard reduction operation, which is efficiently implemented in Thrust. Furthermore, our approach is scalable to multiple GPU devices, as only a very small number of transfers of data between GPU and CPU is required.

We outlined two applications in data analysis which require an efficient calculation of the medians and order statistics. Both applications are easily parallelized for GPUs, and benefit from an efficient GPU-based selection algorithm.

REFERENCES

- [1] I. Al-Furiah, S. Aluru, S. Goil, and S. Ranka. Practical algorithms for selection on coarse-grained parallel computers. *IEEE Trans. on Parallel and Distributed Syst.*, 8:813–824, 1997.
- [2] D. Bader. An improved, randomized algorithm for parallel selection with an experimental study. *J. Parallel Distrib. Comput.*, 64:1051 – 1059, 2004.
- [3] A. Bagirov. A method for minimization of quasidifferentiable functions. *Optimization Methods and Software*, 17:31–60, 2002.
- [4] M. Blum, R.W. Floyd, V. Watt, R.L. Rive, and R.E. Tarjan. Time bounds for selection. *Journal of Computer and System Sciences*, 7:448–461, 1973.
- [5] P.S. Bullen. *Handbook of Means and Their Inequalities*. Kluwer, Dordrecht, 2003.
- [6] T. Calvo and G. Beliakov. Aggregation functions based on penalties. *Fuzzy Sets and Systems*, 161:1420–1436, 2010.
- [7] T. Calvo, R. Mesiar, and R. Yager. Quantitative weights and aggregation. *IEEE Trans. on Fuzzy Systems*, 12:62–69, 2004.
- [8] D. Cederman and P. Tsigas. GPU-Quicksort: A practical quicksort algorithm for graphics processors. *ACM Journal of Experimental Algorithmics*, 14:1.4.1–1.4.24, 2009.
- [9] F.H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley, New York, 1983.
- [10] V.F. Demyanov and A.M. Rubinov. *Constructive Nonsmooth Analysis*. Peter Lang, Frankfurt am Main, 1995.
- [11] C. Gini. *Le Medie*. Unione Tipografico-Editorial Torinese, Milan (Russian translation, Srednie Velichiny, Statistica, Moscow, 1970), 1958.
- [12] N. K. Govindaraju, J. Gray, R. Kumar, and D. Manocha. GPUtera-Sort: High performance graphics coprocessor sorting for large database management. In *Proc. 2006 ACM SIGMOD Intl Conf. on Management of Data*, pages 325–336, 2006.
- [13] N. K. Govindaraju, B. Lloyd, W. Wang, M. Lin, and D. Manocha. Fast computation of database operations using graphic processors. In *Proc. 2004 ACM SIGMOD Intl Conf. on Management of Data*, pages 215 – 226, 2004.
- [14] S. Le Grand. Broad-phase collision detection with CUDA. In H. Nguyen, editor, *GPU Gems 3*, pages 697–721. Addison-Wesley Professional, 2007.
- [15] J. Hoberock and N. Bell. Thrust: A parallel template library, <http://www.meganeurons.com/>, 2010. Version 1.3.0.
- [16] D. Jackson. Note on the median of a set of numbers. *Bulletin of the American Math. Soc.*, 27:160–164, 1921.

- [17] J.E. Kelley. The cutting-plane method for solving convex programs. *J. of SIAM*, 8:703–712, 1960.
- [18] N. Leischner, V. Osipov, and P. Sanders. GPU samplesort. In *IEEE International Parallel and Distributed Processing Symposium*, Atlanta, 2010. IEEE, DOI: 10.1109/IPDPS.2010.5470444.
- [19] R. Maronna, R. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.
- [20] D. Merrill and A. Grimshaw. Revisiting sorting for GPGPU stream architectures. Technical Report CS2010-03, University of Virginia, Department of Computer Science, Charlottesville, VA, USA, 2010.
- [21] NVIDIA. http://developer.download.nvidia.com/compute/cuda/1_1/website/data-parallel_algorithms.html, accessed 1 Feb, 2011.
- [22] NVIDIA. http://developer.nvidia.com/object/cuda_4_0_rc_downloads.html, accessed 20 March, 2011.
- [23] NVIDIA. Tesla datasheet, http://www.nvidia.com/docs/io/43395/nv_ds_tesla_psc_us_nov08_lowres.pdf, accessed 1 December, 2010.
- [24] A.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, 2002.
- [25] P.J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79:871–880, 1984.
- [26] P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.*, 88:1273–1283, 1993.
- [27] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 2003.
- [28] P.J. Rousseeuw and K. Van Driessen. Computing lts regression for large data sets. *Data Mining and Knowledge Discovery*, 12:29–45, 2006.
- [29] N. Satish, M. Harris, and M. Garland. Designing efficient sorting algorithms for manycore GPUs. In *In Proc. of IEEE Intl Parallel and Distributed Processing Symposium (IPDPS 2009), Rome*, DOI: 10.1109/IPDPS.2009.5161005, 2009.
- [30] E.L. Saukas and S.W. Song. Efficient selection algorithms on distributed memory computers. In *ACM/IEEE Conference on Supercomputing*, pages 1–26, San Jose, CA, 1998. IEEE.
- [31] R. Sedgewick. *Algorithms*. Addison-Wesley, Reading, MA, 2nd edition, 1988.
- [32] S. Sengupta, M. Harris, Y. Zhang, and J. D. Owens. Scan primitives for GPU computing. *Graphics Hardware*, pages 97–106, 2007.
- [33] E. Sintorn and U. Assarsson. Fast parallel GPU-sorting using a hybrid algorithm. *J. of Parallel and Distributed Computing*, 68:1381–1388, 2008.
- [34] A. Tiskin. Parallel selection by regular sampling. *LNCS*, 6272:393–399, 2009.
- [35] R. Yager and A. Rybalov. Understanding the median as a fusion operator. *Int. J. General Syst.*, 26:239–263, 1997.

```

typedef thrust::pair<double , unsigned int> Mypair;

template <typename T, typename T1>
struct abs_diff : public thrust::unary_function<T,T1>
{
    const double y;
    abs_diff(double _a) : y(_a) {}
    __host__ __device__
    T1 operator()(const T& a) {
        if(y >= a) return Mypair(y-a,1);
        else return Mypair(a-y,0);
    }
};

template <typename T>
struct plusplus : public thrust::binary_function<T,T,T>
{
    __host__ __device__
    T operator()(const T& a, const T& b)
    { return Mypair(a.first+b.first,a.second+b.second); }
};

void Objective(int n, double t, double * f, double * df)
{
    /* calculates the values of the objective and its subgradient */
    abs_diff<double , Mypair> unary_op(t);
    plusplus<Mypair> binary_op;

    Mypair initpair(0.0,0);
    Mypair result =
        thrust::transform_reduce(data.begin(), data.end(),
                                unary_op, initpair, binary_op);
    *df = 2.0*result.second-n;
    *f = result.first;
}

void SortZ(double * result, double L, double R, int index)
/* Copies the data satisfying L<data[i]<R into Z and returns the n/2-index
order statistic of Z after sorting. */
{
    inside_interval pred(L, R);
    DoubleIterator endZ =
        thrust::copy_if(data.begin(),data.end(),Z.begin(),pred);

    thrust::sort(Z.begin(),endZ);
    *result=Z[n/2-index];
}

```

Fig. 1: A fragment of C code for evaluation of the objective $f(t)$ and its subgradient $g(t)$ in parallel by using Thrust transform_reduce function. The function Objective is called by Algorithm 1, and once it terminates after *maxit* iterations, the function SortZ calculates the median of x by sorting array z containing the data in the pivot interval $[L, R]$.

TABLE I: The mean time (ms) for each method used to calculate the median: data type *float*. The best timing is in boldface.

Method	Size of the data set n							
	8192	32768	131072	524288	2097152	8388608	33554432	134×10^6
Radix Sort (on GPU)	0.27	0.32	0.94	1.93	4.87	17.07	65.68	282.10
Quickselect (on CPU)	0.21	0.93	3.75	14.62	57.49	235.85	940.1	
including:								
- copy to CPU	0.1	0.3	1.1	4	14.5	58.5	232	
- algorithm	0.11	0.63	2.65	10.62	42.99	177.35	708.1	
Quickselect (on GPU)	4.88	21.00	80.26	326.59	1311.60	5530	21951	
Cutting Plane (total)	0.70	0.80	1.04	3.29	4.32	8.34	24.48	88.00
including:								
- CP iterations	0.29	0.32	0.49	2.40	2.90	4.80	14.20	62.00
- copy_if	0.19	0.22	0.25	0.47	0.61	1.64	5.98	13.90
- Radix sort of \mathbf{z}	0.22	0.26	0.30	0.42	0.81	1.90	4.30	12.10
Bisection	0.77	1.11	1.44	3.44	10.7	51.6	216.4	
Brent's minimization	0.97	1.12	1.4	2.25	5.85	20.75	86.4	
Brent's nonlinear eqn	0.68	0.79	0.87	2.34	3.92	8.79	31.25	

TABLE II: The mean time (ms) for each method used to calculate the median: data type *double*. The best timing is in boldface.

Method	Size of the data set n							
	8192	32768	131072	524288	2097152	8388608	33554432	134×10^6
Radix Sort (on GPU)	0.45	0.59	1.98	4.65	15.17	57.52	226.39	820.76
Quickselect (on CPU)	0.29	1.12	4.52	17.12	68.11	280.83	1107.82	
including:								
- copy to CPU	0.1	0.48	1.74	6.1	19.1	64	455	
- algorithm	0.19	0.64	2.78	11.02	49.01	216.83	652.8	
Quickselect (on GPU)	5.25	20.72	87.10	358.58	1429.4	6175.6	24056.7	
Cutting Plane (total)	1.00	1.10	3.24	3.68	5.26	11.95	37.84	139.84
including:								
- CP iterations	0.44	0.5	2.10	2.36	3.20	6.80	23.00	94.00
- copy_if	0.28	0.29	0.72	0.83	0.96	2.95	9.74	22.84
- Radix sort of \mathbf{z}	0.28	0.31	0.42	0.49	1.10	2.20	5.10	23.00
Bisection	1.02	1.24	4.12	5.41	12.1	61.2	298	
Brent's minimization	1.01	1.19	3.12	4.02	5.45	26.75	92.1	
Brent's nonlinear eqn	0.97	1.04	2.81	3.25	4.49	11.14	39.2	

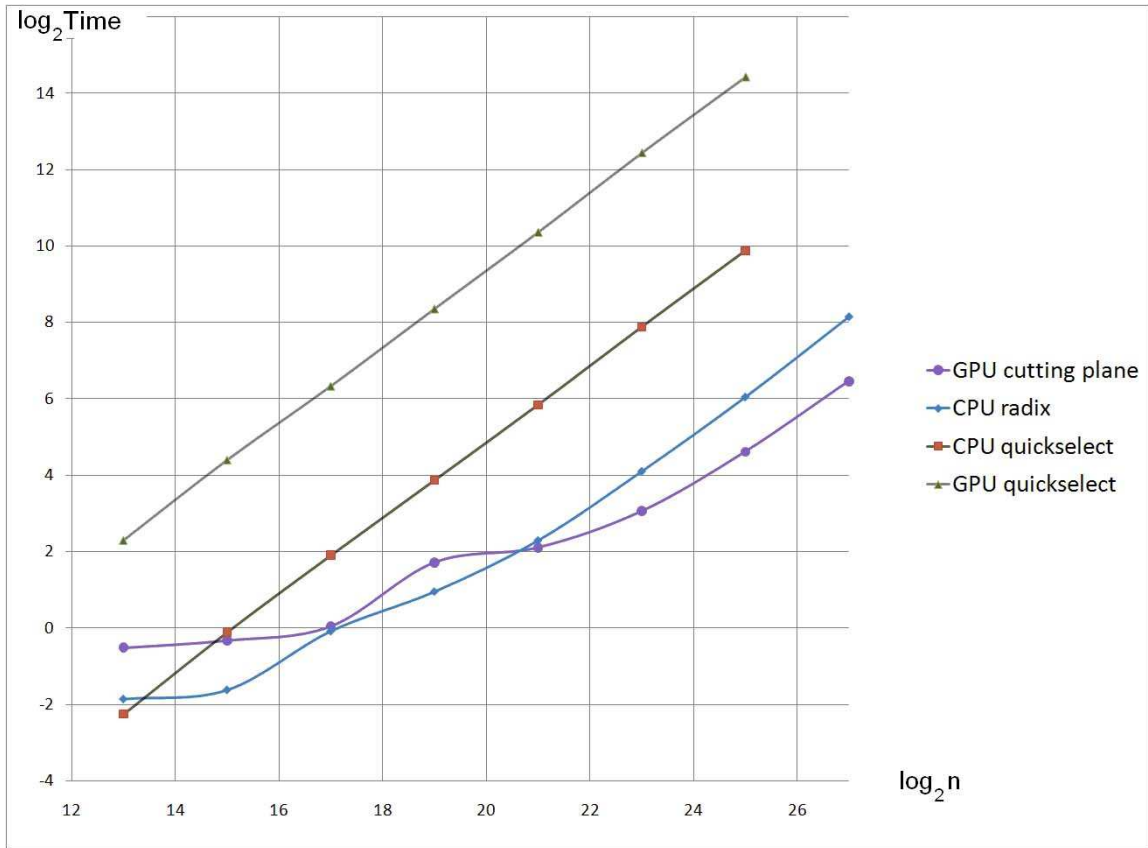


Fig. 2: Average time (in milliseconds, on a logarithmic scale) used by different methods to calculate the median, for data type `float`.

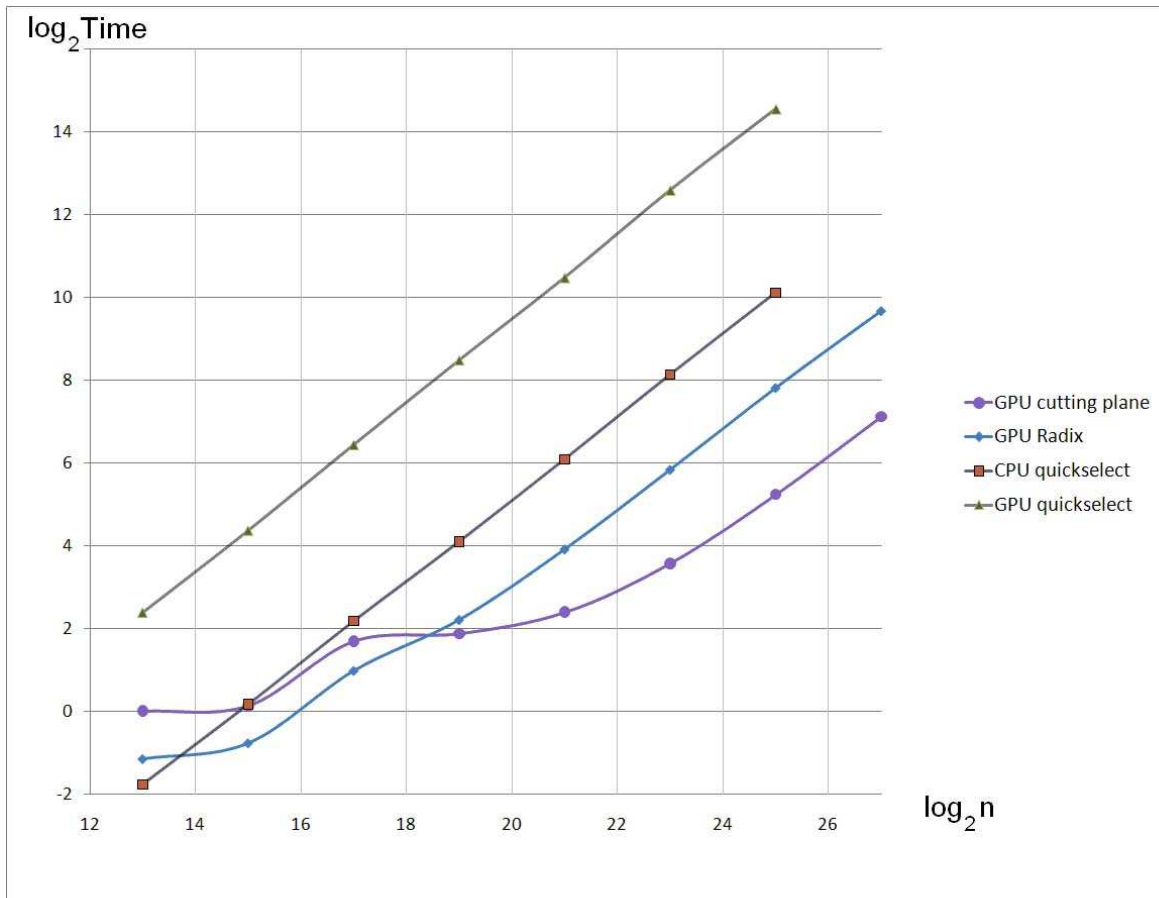


Fig. 3: Average time (in milliseconds, on a logarithmic scale) used by different methods to calculate the median, for data type double.

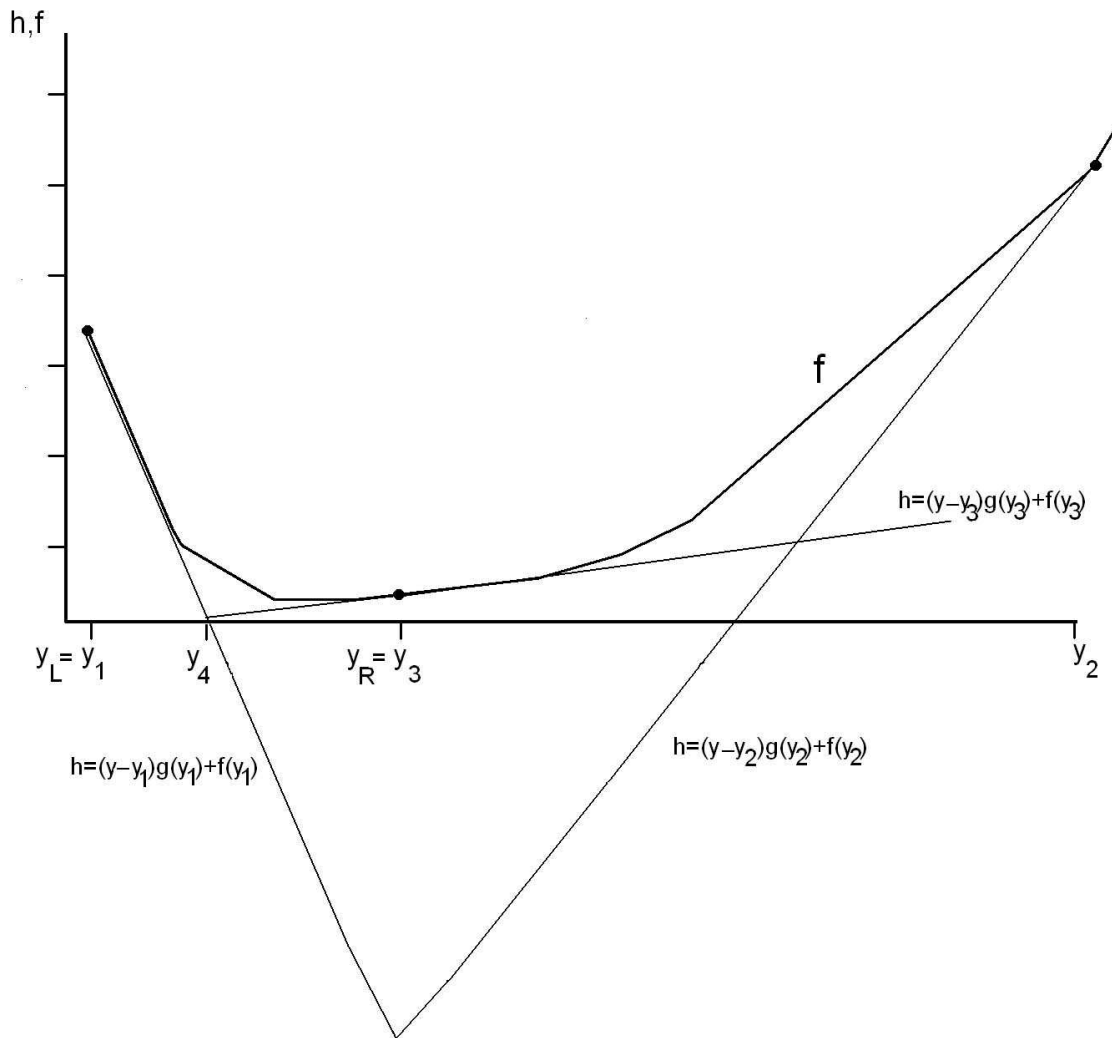


Fig. 4: Illustration of the cutting plane algorithm. The thick solid line is the graph of the objective f . The minimizer of f is always bracketed by y_L and y_R , which are updated at every iteration. The point of intersection of two lines tangent to the graph of f yields the value y_{k+1} of the new iteration.

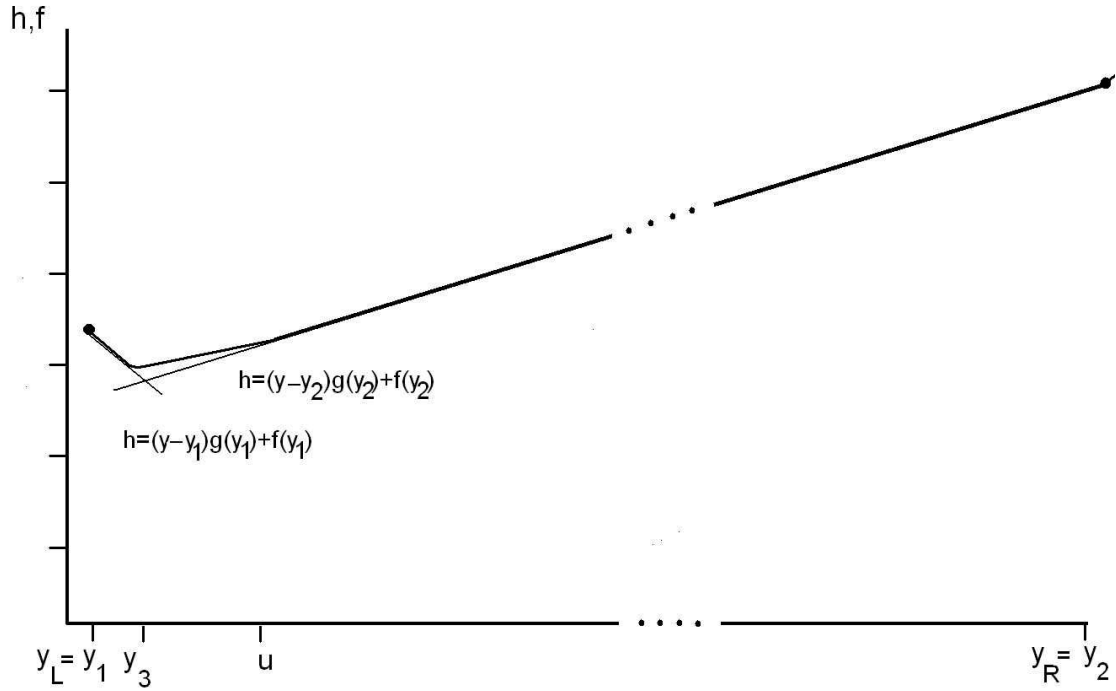


Fig. 5: Non-sensitivity of the cutting plane algorithm to extreme values of x_i . Even if the right end of the interval $[y_L, y_R]$ is very large, the cutting plane algorithm avoids exploring the uninteresting interval $[u, y_R]$, because on that interval f coincides with $h_2(y) = (y - y_R)g(y_R) + f(y_R)$, and the next iteration y_3 cannot be selected in $[u, y_R]$. In contrast, bisection makes many iterations in $[u, y_R]$, the larger y_R , the more iterations in $[u, y_R]$. Similarly, Brent's method attempts parabolic fit using 3 points from $[u, y_R]$, but since these points are collinear, it reverts to golden section, which again explores $[u, y_R]$.